Exploring Memorization in Adversarial Training

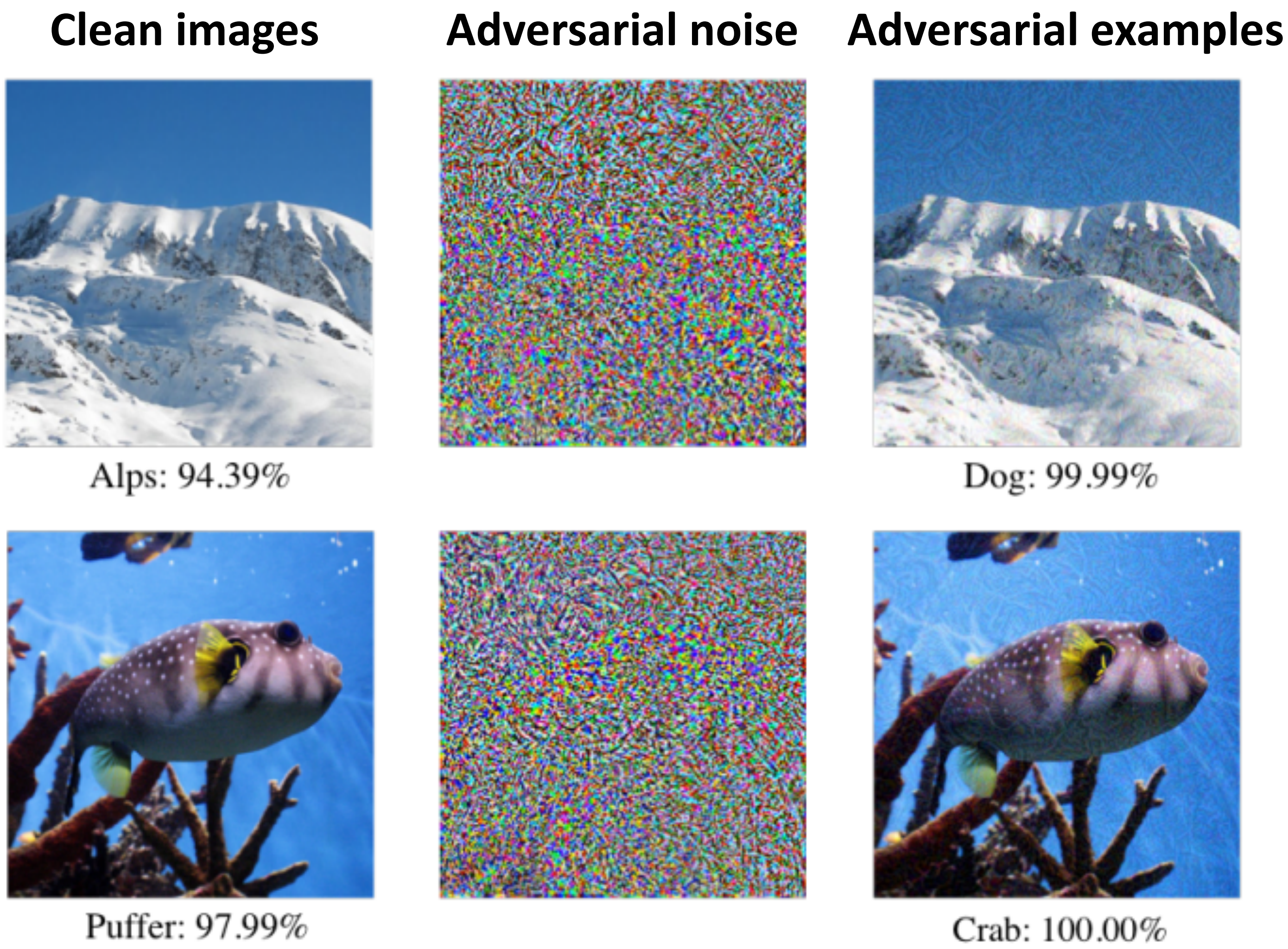
Yinpeng Dong[1,2], Ke Xu[3], Xiao Yang[1], Tianyu Pang[1], Zhijie Deng[1], Hang Su[1], Jun Zhu[1,2]

[1] Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua University [2] RealAI [3] CMU

{dongyinpeng, suhangss, dcszj}@mail.tsinghua.edu.cn; kx1@andrew.cmu.edu;

{yangxiao19, pty17, dzj17}@mails.tsinghua.edu.cn

Adversarial Examples



(Figure is from Dong et al. 2018)

Adversarial Training

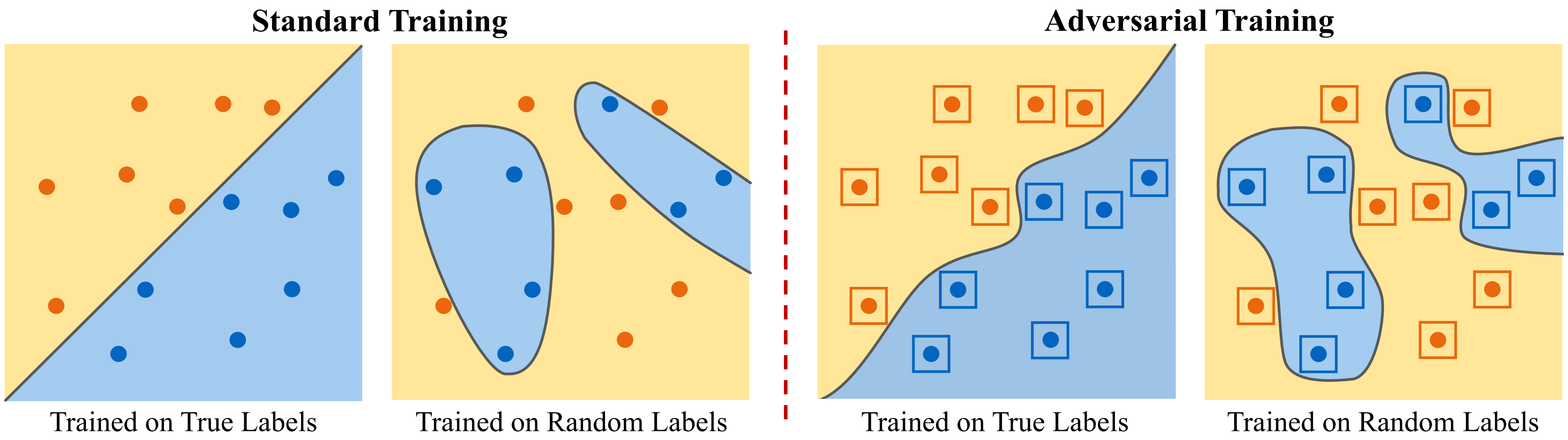
- From the optimization view, adversarial training (AT) can be formulated as a minimax optimization problem (Madry et al., 2018)

$$\overbrace{\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\max_{\delta_i \in S} L(f_\theta(x_i + \delta_i), y_i)}_{\text{Inner maximization: generate an adversarial example}}}^{\text{Outer minimization: train a robust classifier}} \qquad S = \{\delta : \|\delta\|_\infty \le \epsilon\}$$

- Solve the inner maximization by projected gradient descent

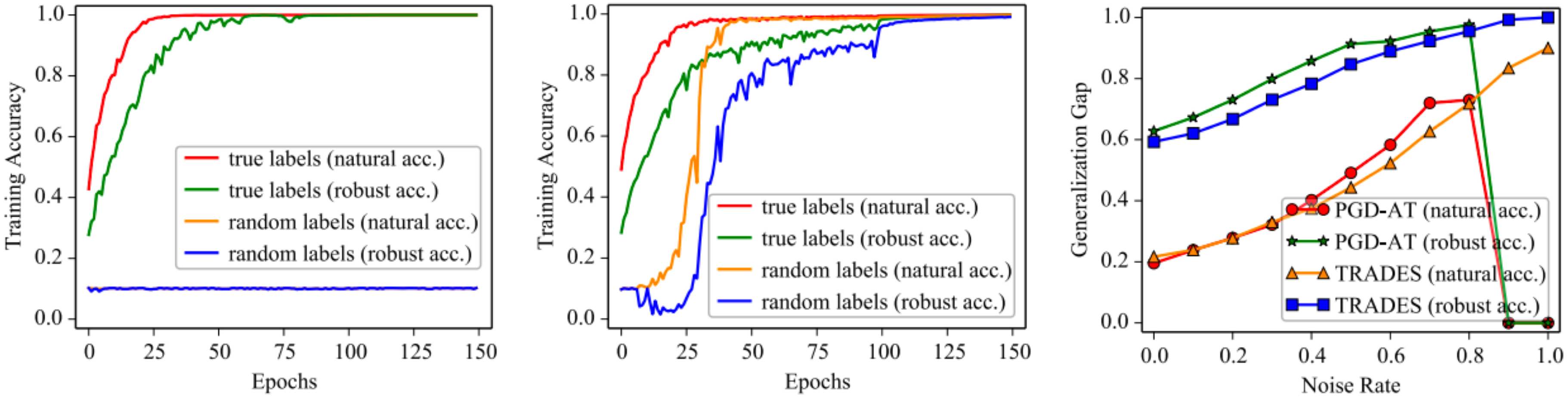
$$\delta_i^{t+1} = \Pi_S\left(\delta_i^t + \alpha \cdot \mathrm{sign}\left(\nabla_x L\left(f_\theta\left(x_i + \delta_i^t\right), y_i\right)\right)\right)$$

Memorization



Goal: To facilitate a deeper understanding of model capacity, training convergence, robust generalization, and robust overfitting of the adversarially trained models.

AT with random labels



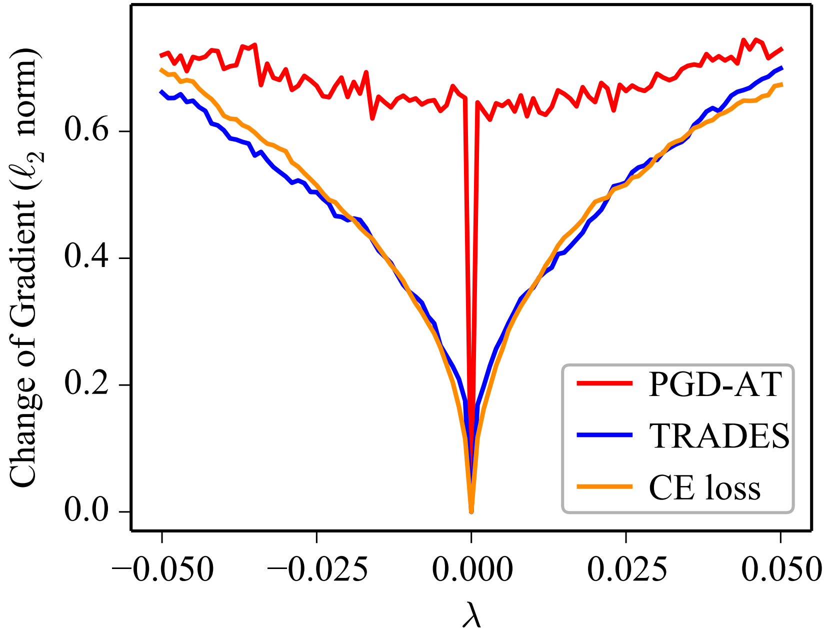
(a) PGD-AT (b) TRADES (c) Generalization gap growth

Finding: PGD-AT cannot converge with random labels, but TRADES can.

- This finding holds with different training settings, including network architecture, attack steps, optimizer, perturbation budget, regularizations

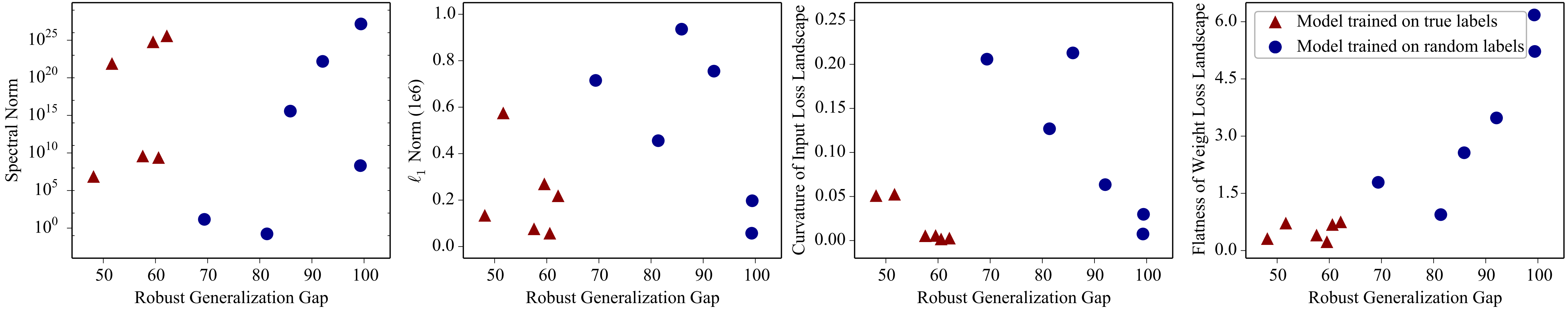
Convergence Analysis

- **Gradient instability issue**: the gradient of the adversarial loss in PGD-AT changes more abruptly than TRADES.



We measure the L2 distance between the gradient at θ and $\theta + \lambda d$

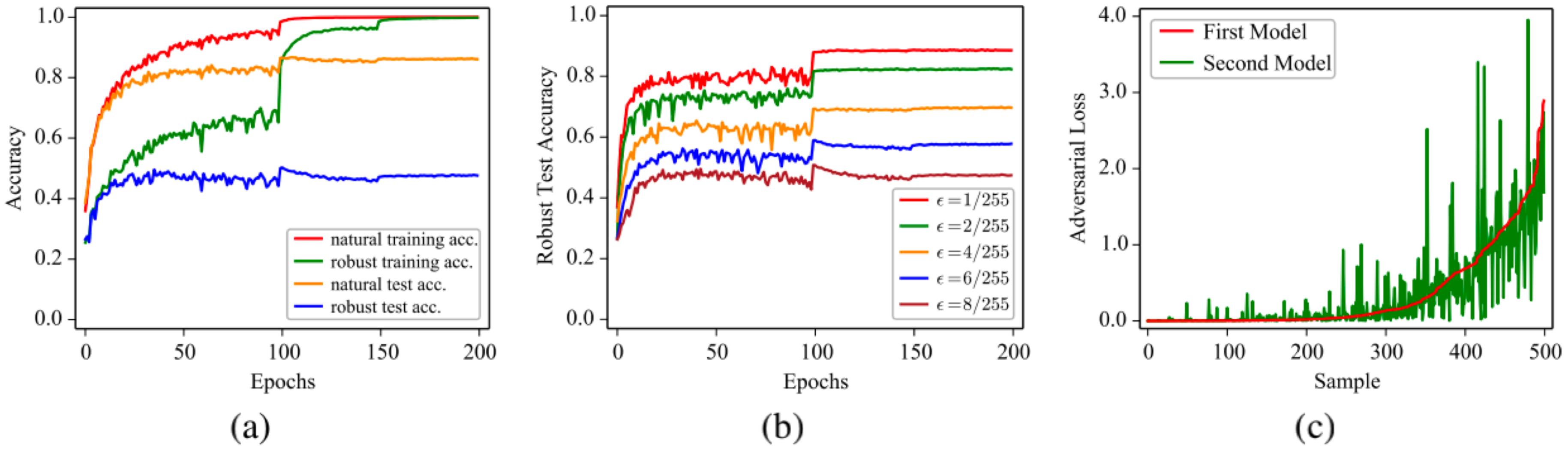
Generalization Analysis



- We consider two norm-based measures and two sharpness/flatness-based measures.
- **Finding**: None of them can adequately explain and ensure robust generalization.

Robust Overfitting Analysis

- **Argument: robust overfitting is caused by excessive memorization of (noisy) one-hot labels.**



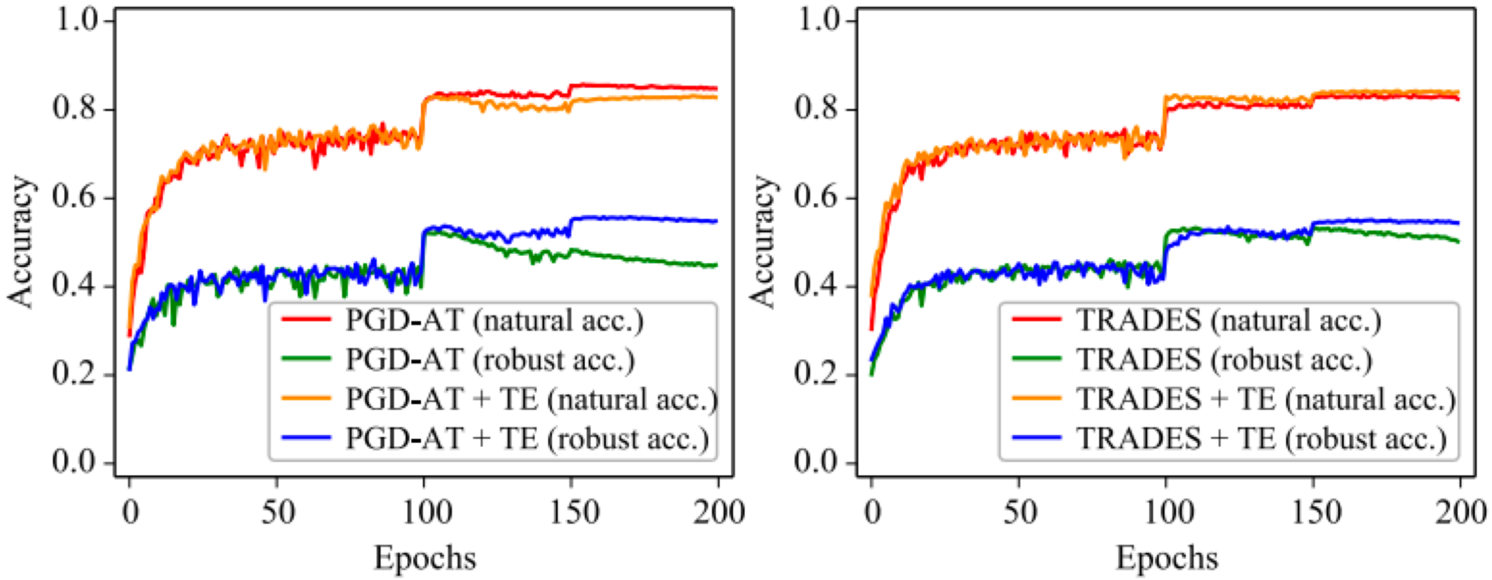
1. Robust overfitting does not occur with a smaller perturbation budget (e.g., $\epsilon = 1/255$).
2. The hard examples are consistent across different networks.

Mitigation Algorithm

- Incorporate the **Temporal Ensembling (TE)** approach into AT

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{\delta_i \in S} \{ L(f_\theta(x_i + \delta_i), y_i) + w \cdot \| f_\theta(x_i + \delta_i) - \hat{p}_i \|_2^2 \}$$

- $\hat{p}_i = \frac{p_i}{\|p_i\|_2}, p_i = \eta p_i + (1 - \eta) f_\theta(x_i)$



Learning curves on CIFAR-10

Empirical Results

Method	Natural Accuracy			PGD-10			PGD-1000			C&W-1000			AutoAttack		
	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff
PGD-AT	**83.75**	**84.82**	-1.07	52.64	44.92	7.72	51.22	42.74	8.48	50.11	43.63	7.48	47.74	41.84	5.90
PGD-AT+TE	82.35	82.79	**-0.44**	**55.79**	**54.83**	**0.96**	**54.65**	**53.30**	**1.35**	**52.30**	**51.73**	**0.57**	**50.59**	**49.62**	**0.97**
TRADES	81.19	82.48	-1.29	53.32	50.25	3.07	52.44	48.67	3.77	49.88	48.14	1.74	49.03	46.80	2.23
TRADES+TE	**83.86**	**83.97**	**-0.11**	**55.15**	**54.42**	**0.73**	**53.74**	**53.03**	**0.71**	**50.77**	**50.63**	**0.14**	**49.77**	**49.20**	**0.57**

(a) The evaluation results on **CIFAR-10**.

Method	Natural Accuracy			PGD-10			PGD-1000			C&W-1000			AutoAttack		
	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff
PGD-AT	**57.54**	**57.51**	**0.03**	29.40	21.75	7.65	28.54	20.63	7.91	27.06	21.17	5.89	24.72	19.34	5.38
PGD-AT+TE	56.45	57.12	-0.67	**31.74**	**30.24**	**1.50**	**31.27**	**29.80**	**1.47**	**28.27**	**27.36**	**0.91**	**26.30**	**25.34**	**0.96**
TRADES	57.98	56.32	1.66	29.93	27.70	2.23	29.51	26.93	2.58	25.46	24.42	1.04	24.61	23.40	1.21
TRADES+TE	**59.35**	**58.72**	**0.63**	**31.09**	**30.12**	**0.97**	**30.54**	**29.45**	**1.09**	**26.61**	**25.94**	**0.67**	**25.27**	**24.55**	**0.72**

(b) The evaluation results on **CIFAR-100**.

Method	Natural Accuracy			PGD-10			PGD-1000			C&W-1000			AutoAttack		
	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff	Best	Final	Diff
PGD-AT	89.00	90.55	-1.55	54.51	46.97	7.54	52.22	42.85	9.37	48.66	44.13	4.53	46.61	38.24	8.37
PGD-AT+TE	**90.09**	**90.91**	**-0.82**	**59.74**	**59.05**	**0.69**	**57.71**	**56.46**	**1.25**	**54.55**	**53.94**	**0.61**	**51.44**	**50.61**	**0.83**
TRADES	**90.88**	**91.30**	**-0.42**	59.50	57.04	2.46	52.78	50.17	2.61	52.76	50.53	2.23	40.36	38.88	1.48
TRADES+TE	89.01	88.52	0.49	**59.81**	**58.49**	**1.32**	**58.24**	**56.66**	**1.58**	**54.00**	**53.24**	**0.76**	**51.45**	**50.16**	**1.29**

(c) The evaluation results on **SVHN**.

Thanks for listening

Code is available at:
https://github.com/dongyp13/memorization-AT